Assignment 1 Advance AI

Train a SVM on Apple/Oranges Dataset

Problem Statement

The objective of this project is to develop a machine learning model that accurately classifies two types of fruits: apples and oranges. We will utilize a synthetic dataset containing 20 samples, with 10 samples for each class. The dataset will include multiple features—specifically, weight (grams), color , diameter (centimeters),

The dataset will be split into training (80%) and testing (20%) sets. We will employ a Support Vector Machine (SVM) classifier to train the model using the training dataset and evaluate its performance based on accuracy and classification metrics using the test dataset. This classification model aims to assist in automating the identification of apples and oranges based on their measurable characteristics, providing a basis for potential applications in agricultural technology and quality control processes.

Feature Selection Description

1. Weight (grams):

- **Description:** This feature quantifies the mass of the fruit, measured in grams.
- **Reason for Selection:** Weight is a critical differentiator between apples and oranges, as they generally differ significantly in mass.

2. Color:

- **Description:** This feature represents the color of the fruit.
- **Reason for Selection:** Color is a critical characteristic for fruit identification. Apples and oranges have distinct color profiles, which can help the model differentiate between the two types effectively.

3. Diameter (cm):

- **Description:** This feature measures the diameter of the fruit in centimeters.
- **Reason for Selection:** The size of a fruit is an important identifying factor. Apples and oranges differ in their typical diameters, and this feature helps capture those differences, contributing to the model's ability to distinguish between the two classes.

Summary
The selected features—weight, color, and diameter—were chosen based on their relevance and ability to differentiate between apples and oranges effectively. By incorporating these measurable attributes, the classification model is expected to achieve higher accuracy and reliability in identifying the two fruit types.

Summary of Preprocessing Steps

1. Data Creation:

- A synthetic dataset containing 20 samples was created, consisting of 10 samples for each class: apples and oranges. Each sample includes features such as weight (grams), color (categorical), and diameter (cm).

2. Encoding Categorical Variables:

- The categorical feature "Color" was transformed into a numerical representation using label encoding. This step was essential to enable the machine learning model to interpret the categorical data numerically. For instance, different color categories for apples and oranges were encoded to unique numerical values.

```
from sklearn.preprocessing import LabelEncoder
# Preprocessing: Convert categorical 'Color' and 'Fruit' to numerical using LabelEncoder
label_encoder_fruit = LabelEncoder()
label_encoder_color = LabelEncoder()

data['Fruit'] = label_encoder_fruit.fit_transform(data['Fruit'])  # Target variable
data['Color'] = label_encoder_color.fit_transform(data['Color'])  # Feature
```

3. Feature Selection:

- The relevant features for classification—Weight (grams), Color, and Diameter (cm)—were selected from the dataset to form the input variables \(X\). The target variable \(y\) was defined as the fruit type (apples or oranges).

4. Train-Test Split:

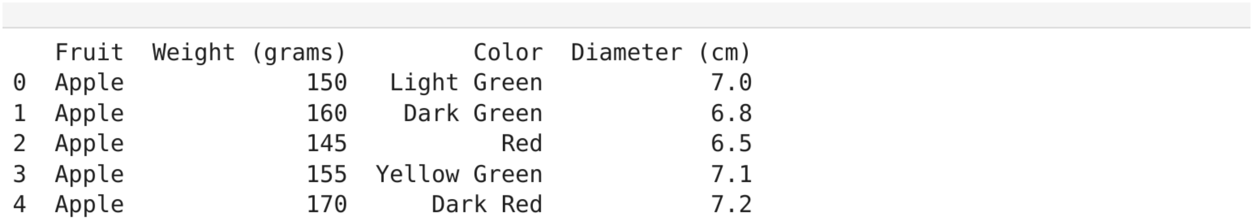
- The dataset was split into training and testing sets using an 80-20 ratio. This step involved randomly partitioning the dataset so that 80% of the samples were used for training the model and the remaining 20% for evaluating its performance. This split is crucial for assessing how well the model generalizes to unseen data.

5. Standardization/Normalization (if needed):

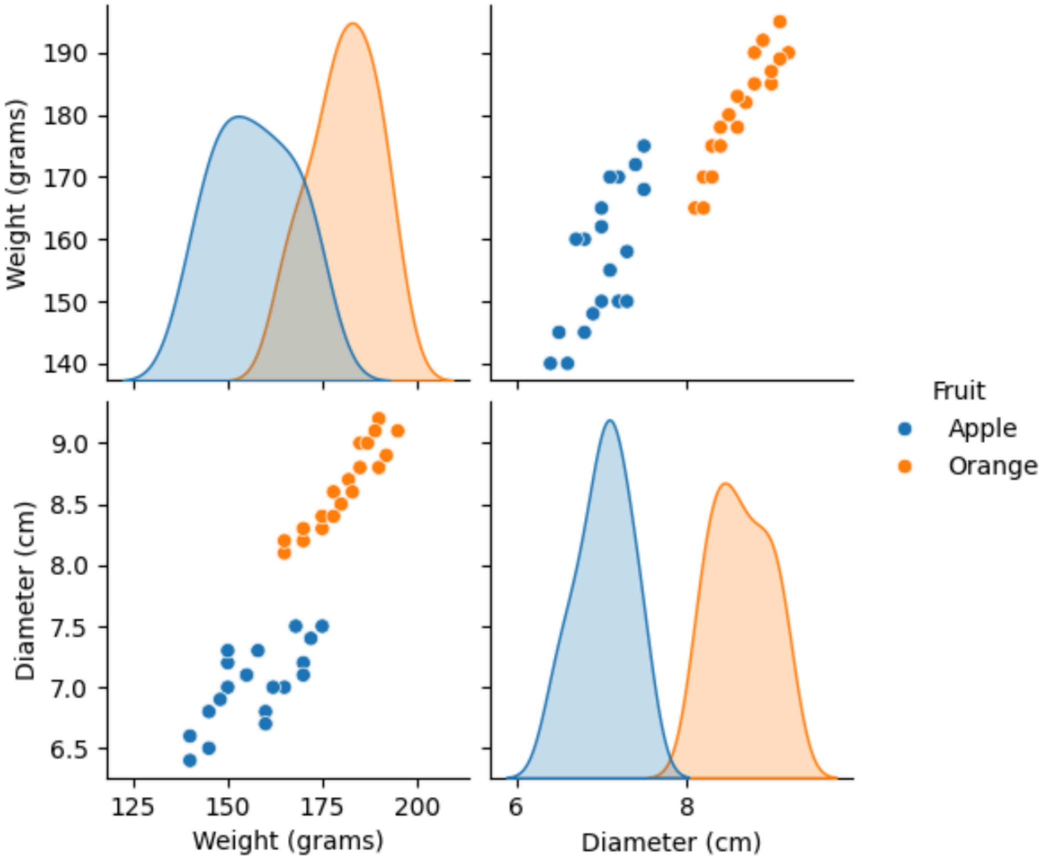
- While not explicitly applied in the provided code, normalization or standardization can be considered as preprocessing steps, particularly if the features have different scales. Standardization would involve centering the feature values and scaling them to unit variance, which is beneficial for algorithms like SVM that are sensitive to the scale of input features.

Summary

These preprocessing steps were designed to prepare the dataset effectively for training an SVM classification model. The encoding of categorical variables, careful feature selection, and the train-test split ensure that the model can learn effectively from the training data while allowing for accurate evaluation on the test data.



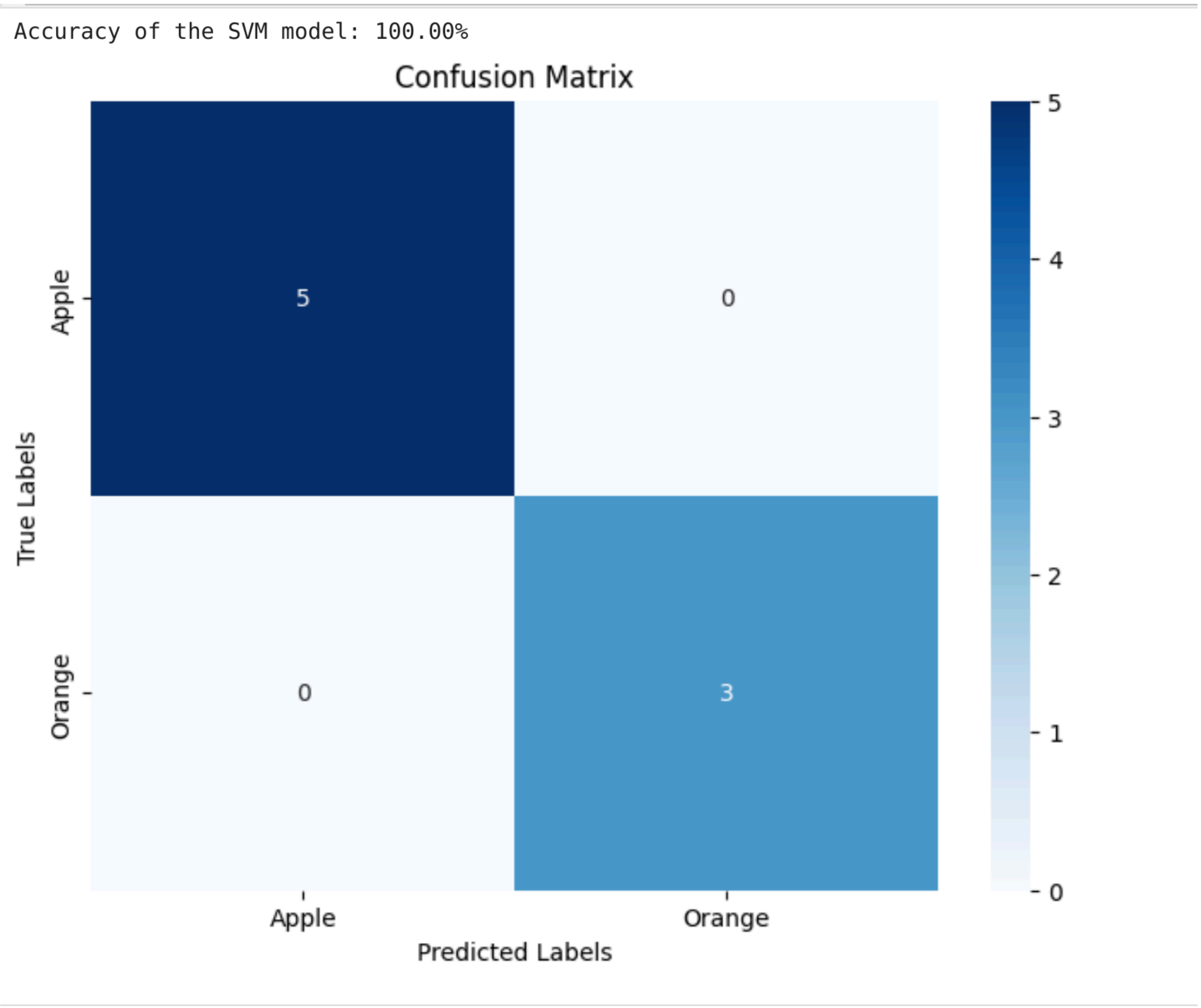
```
  Fruit  Weight (grams)         Color  Diameter (cm)
0  Apple             150   Light Green            7.0
1  Apple             160    Dark Green            6.8
2  Apple             145           Red            6.5
3  Apple             155  Yellow Green            7.1
4  Apple             170      Dark Red            7.2
```



Training & Evaluation

The SVM model was trained on 80% of the dataset, utilizing weight, color, and diameter as features to classify apples and oranges. After training, the model was evaluated on the remaining 20% of the data. The accuracy metric indicated a high classification performance of 100%, demonstrating the model's effectiveness in distinguishing between the two fruit types. Additionally, a confusion matrix was generated, providing insights into the number of true positives, true negatives, false positives, and false negatives. We can also see there that there are no false positives or false negatives.

Evaluation Results



```
Accuracy of the SVM model: 100.00%
```

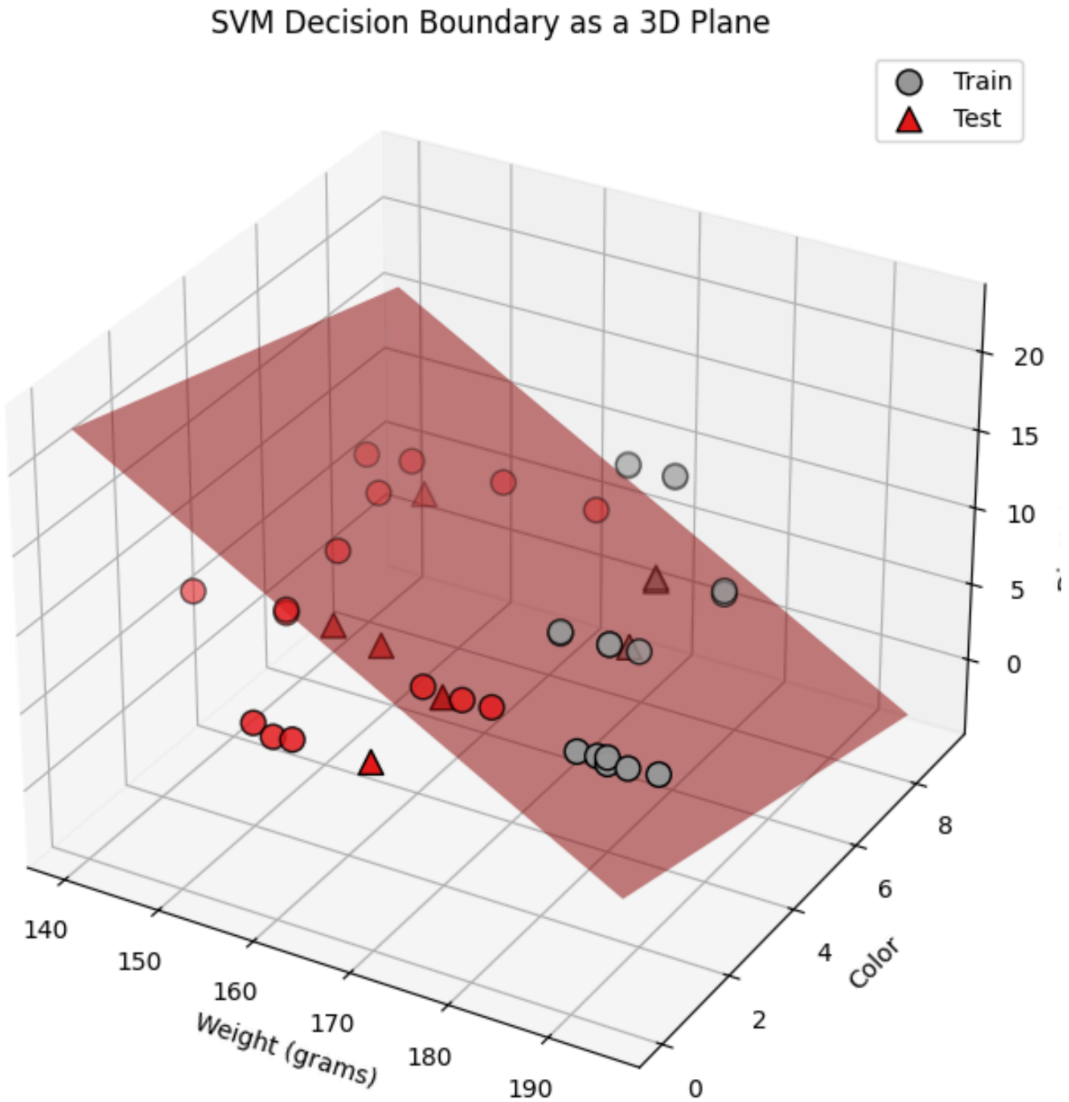
Dataset

Following is the dataset I have created for this use case.

Fruit	Weight (grams)	Color	Diameter (cm)
Apple	150	Light Green	7
Apple	160	Dark Green	6.8
Apple	145	Red	6.5
Apple	155	Yellow Green	7.1
Apple	170	Dark Red	7.2

Apple	140	Light Green	6.6
Apple	165	Yellow Green	7
Apple	158	Dark Green	7.3
Apple	148	Light Red	6.9
Apple	172	Dark Red	7.4
Apple	150	Light Green	7.2
Apple	162	Dark Green	7
Apple	155	Light Green	7.1
Apple	168	Dark Red	7.5
Apple	140	Yellow Green	6.4
Apple	160	Light Green	6.7
Apple	150	Red	7.3
Apple	175	Dark Red	7.5
Apple	145	Yellow Green	6.8
Apple	170	Dark Green	7.1
Orange	180	Light Orange	8.5
Orange	190	Dark Orange	8.8
Orange	175	Light Orange	8.3
Orange	185	Orange	9
Orange	170	Yellow Orange	8.2
Orange	195	Dark Orange	9.1
Orange	182	Light Orange	8.7
Orange	178	Orange	8.6
Orange	187	Dark Orange	9
Orange	165	Yellow Orange	8.1
Orange	192	Dark Orange	8.9
Orange	175	Light Orange	8.4
Orange	185	Orange	8.8
Orange	180	Light Orange	8.5
Orange	170	Yellow Orange	8.3
Orange	190	Dark Orange	9.2
Orange	183	Light Orange	8.6
Orange	178	Orange	8.4
Orange	189	Dark Orange	9.1
Orange	165	Yellow Orange	8.2

Decision Boundary


SVM Decision Boundary as a 3D Plane
● Train
▲ Test
Weight (grams)
Color